View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Digital Repository @ Iowa State University

IOWA STATE UNIVERSITY
Digital Repository

Statistics Publications | Statistics

12-2005

Hot Deck Imputation for the Response Model

Wayne A. Fuller
Iowa State University, waf@iastate.edu

Jae Kwang Kim
Yonsei University, jkim@iastate.edu

Follow this and additional works at: https://lib.dr.iastate.edu/stat_las_pubs

 Part of the Design of Experiments and Sample Surveys Commons, Probability Commons, Statistical Methodology Commons, and the Statistical Models Commons

The complete bibliographic information for this item can be found at https://lib.dr.iastate.edu/stat_las_pubs/312. For information on how to cite this item, please visit http://lib.dr.iastate.edu/howtocite.html.

This Article is brought to you for free and open access by the Statistics at Iowa State University Digital Repository. It has been accepted for inclusion in Statistics Publications by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.

Hot Deck Imputation for the Response Model

Abstract

Hot deck imputation is a procedure in which missing items are replaced with values from respondents. A model supporting such procedures is the model in which response probabilities are assumed equal within imputation cells. An efficient version of hot deck imputation is described for the cell response model and a computationally efficient variance estimator is given. An approximation to the fully efficient procedure in which a small number of values are imputed for each nonrespondent is described. Variance estimation procedures are illustrated in a Monte Carlo study.

Keywords

Nonresponse, Fractional imputation, Response probability, Replication variance estimation

Disciplines

Design of Experiments and Sample Surveys | Probability | Statistical Methodology | Statistical Models

Comments

This article is published as Fuller, W.A. and Kim, J.K. (2005). Hot Deck Imputation for the Response Model, *Survey Methodology* 31, 139-149. Posted with permission.

This article is available at Iowa State University Digital Repository: https://lib.dr.iastate.edu/stat_las_pubs/312

Vol. 31, No. 2, pp. 139-149
Statistics Canada, Catalogue No. 12-001

Hot Deck Imputation for the Response Model

Wayne A. Fuller and Jae Kwang Kim [1]

Abstract

Hot deck imputation is a procedure in which missing items are replaced with values from respondents. A model supporting such procedures is the model in which response probabilities are assumed equal within imputation cells. An efficient version of hot deck imputation is described for the cell response model and a computationally efficient variance estimator is given. An approximation to the fully efficient procedure in which a small number of values are imputed for each nonrespondent is described. Variance estimation procedures are illustrated in a Monte Carlo study.

Key Words: Nonresponse; Fractional imputation; Response probability; Replication variance estimation.

1. Introduction

Imputation is used in sample surveys as a method of handling item nonresponse. In hot deck imputation, the imputed values are functions of the respondents in the current sample. Sande (1983) and Ford (1983) contain descriptions of hot deck imputation. Kalton and Kasprzyk (1986) and Little and Rubin (2002) review various imputation procedures.

In one version of hot deck imputation, the imputed value is the value of a respondent in the same imputation cell, where the imputation cells form an exhaustive and mutually exclusive subdivision of the population. In random hot deck imputation, respondents are assigned values at random from respondents in the same imputation cell. The record providing the value is called the *donor* and the record with the missing value is called the *recipient*.

The variance of the imputed estimator is generally larger than the complete sample variance because nonresponse reduces sample size and because the imputed estimator may contain a component due to random imputation. Rao and Shao (1992) proposed an adjusted jackknife method for hot-deck imputation where the first phase units are selected with-replacement. Rao and Sitter (1995) discussed the adjusted jackknife variance estimation method for ratio imputation. Rao (1996) and Sitter (1997) applied the adjusted jackknife method to regression imputation. Shao, Chen and Chen (1998) apply the idea of Rao and Shao (1992) to the balanced repeated replication method. Shao and Steel (1999) propose variance estimation for survey data with composite imputation, where more than one imputation method is used, and the sampling fractions are included in the variance expressions. Yung and Rao (2000) applied the adjusted jackknife method to imputed estimators constructed with a poststratified sample. Rubin (1987) and Rubin and Schenker (1986) suggested multiple imputation procedures. Tollefson and Fuller (1992), and Särndal (1992) proposed imputation methods and corresponding variance estimators. Kim and Fuller (2004) studied the use of fractional imputation for the model in which observations in an imputation cell are independently and identically distributed.

In this paper, we consider hot deck imputation for a population divided into imputation cells. The response model is described in section 2. In section 3, we introduce fully efficient fractional imputation and present a variance estimation method for the imputation estimator, under the assumptions that the probability of nonresponse is constant within a cell. In section 4 we suggest a modification of the fully efficient method that uses a smaller number of donors. In section 5, an example is introduced to illustrate the actual implementation of the proposed method. In section 6, results of a simulation study are reported. Summary is presented in the last section.

2. Basic Setup

Consider a population of N elements identified by a set of indices $U = \{1, 2, \ldots, N\}$. Associated with each unit i in the population there is a study variable y_i and a vector $\mathbf{x}_i$ of auxiliary information. The set of vectors, $(y_i, \mathbf{x}_i)$, $i = 1, 2, \ldots, N$, is denoted by F.

Let A denote the indices of the elements in a sample selected by a set of probability rules called the *sampling mechanism*. Let the population quantity of interest be θ_N, let $\hat{\theta}$ be a full sample, linear-in-y, estimator of θ_N, and write

$$\hat{\theta} = \sum_{i \in A} w_i \, y_i. \tag{1}$$

1. Wayne A. Fuller, Department of Statistics, Iowa State University, Ames, IA, 50011 U.S.A.; Jae Kwang Kim, Department of Applied Statistics, Yonsei University, Seoul, 120-749, Korea.

If w_i is the inverse of the selection probability, then $\hat{\theta}$ is unbiased for the population total.

Let A_R and A_M denote the set of indices of the sample respondents and sample nonrespondents, respectively. Define the response indicator function

$$R_i = \begin{cases} 1 & \text{if } i \in A_R \\ 0 & \text{if } i \in A_M \end{cases} \tag{2}$$

and let $\mathbf{R} = \{(i, R_i); i \in A\}$. The distribution of $\mathbf{R}$ is called the *response mechanism*.

Assume that the finite population U is made up of G imputation cells, where the set of elements in cell g is U_g. Let n_g be the number of sample elements in imputation cell g and let $r_g, r_g > 0,$ be the number of respondents in imputation cell g. Assume the within-cell uniform response model in which the r_g responses in a cell are equivalent to a Poisson sample selected with equal probabilities from the n_g elements.

Fractional imputation is a procedure in which more than one donor is used per recipient. Kalton and Kish (1984) suggested fractional imputation as an efficient imputation procedure. The method was discussed by Fay (1996). Let d_{ij} be the number of times that y_i is used as donor for the missing y_j and define $\mathbf{d} = \{d_{ij}; i \in A_R, j \in A_M\}$. The distribution of $\mathbf{d}$ is called the *imputation mechanism*. Let w_{ij}^* be the factor applied to the original weight for element j when y_i is used as a donor for element j. For element $j, j \in A_M$,

$$Y_{Ij} = \sum_{i \in A_R} w_{ij}^* y_i \tag{3}$$

is the weighted mean of the respondent values. The factor w_{ij}^* is called the *imputation fraction*. It is the fraction that donor i donates for the missing item y_j. Note that $w_{ii}^* = 1$ for $i \in A_R$ and $w_{ij}^* = 0$ for $i \neq j, i, j \in A_R$. The sum of the imputation fractions for a missing item is restricted to equal one,

$$\sum_{i \in A_R} w_{ij}^* = 1, \quad \forall j \in A. \tag{4}$$

An estimator with the imputed values defined in (3) and some $w_{ij}^* < 1$ is called a *fractionally imputed* estimator.

A linear-in-y imputation estimator can be written in the form

$$\hat{\theta}_I = \sum_{i \in A_R} \left(\sum_{j \in A} w_j w_{ij}^* \right) y_i \tag{5}$$

$$=: \sum_{i \in A_R} \alpha_i y_i, \tag{6}$$

where the notation A =: B means that B is defined to be equal to A. The sum of $w_{ij}^* w_j$ over all recipients for which i is a donor (including acting as a donor for itself), denoted by α_i, is the total weight of donor i. If a responding unit i is not used as a donor, except for itself, then $\alpha_i = w_i$.

3. Fully Efficient Fractional Imputation

Assume all elements in an imputation cell have the same probability of responding and assume the responses are independent. Then the overall distribution of an imputed estimator under the response model can be obtained by using the probability structure of multiple phase sampling, where the response model is treated as the second phase sampling mechanism.

If the response probabilities in a cell are uniform, then a reasonable estimator of the total is the weighted sum of ratio estimators

$$\hat{\theta}_{\text{FE}} = \sum_{g=1}^{G} \left(\sum_{i \in A \cap U_g} w_i \right) \frac{\sum_{i \in A_R \cap U_g} w_i y_i}{\sum_{i \in A_R \cap U_g} w_i}. \tag{7}$$

In the context of two phase sampling, Kott and Stukel (1997) call the estimator (7) a reweighted expansion estimator. The estimator (7) is called fully efficient because it contains no variability due to random selection of donors. If the w_i are the same for all elements in a cell, the ratio

$$\left(\sum_{i \in A_R \cap U_g} w_i \right)^{-1} \sum_{i \in A_R \cap U_g} w_i y_i \tag{8}$$

is a simple mean and, hence, unbiased for the cell mean given that there is at least one respondent in the cell. If the w_i in a cell are not equal, then (8) is subject to ratio bias. It is possible for the number of elements in a cell, n_g, to be positive and the number of respondents, r_g, to be zero. If this occurs in practice, cells will be collapsed.

The large sample properties of the estimator can be obtained for a sequence of populations and samples. Assume the population is composed of G_v mutually exclusive and exhausted cells, where v is the index of the sequence. Assume the variance of a full sample estimator of the mean is $O(n_v^{-1})$, where n_v is the size of the sample selected from the v^{th} population. Assume responses are independent. Then, under regularity conditions, the procedures used by Kim, Navarro and Fuller (2005) in the proof of their Theorem 2.1 can be used to show that estimator (7) satisfies

$$\hat{\theta}_{\text{FE}_v} = \hat{\theta}_v + \sum_{g_v=1}^{Gv} \sum_{i \in A_{gv}} w_{iv} (\pi_{gv}^{-1} R_{iv} - 1) e_{iv} + o_p(n_v^{-1/2} N_v), \tag{9}$$

where $e_{iv} = y_{iv} - \bar{Y}_{gv}$, A_{gv} is the set of sample indices in the g^{th} cell for the v^{th} sample, $\bar{Y}_{gv}$ is the population mean of the y–variable in cell gv of population F_v, π_{gv} is the probability that an element in cell gv responds, and F_v denotes the v^{th} population. Also

$$V(\tilde{\theta}_{\text{FE}v} \mid F_v) = V(\hat{\theta}_v \mid F_v) + E\left\{ \sum_{g_v=1}^{Gv} \pi_{gv}^{-1}(1-\pi_{gv}) \sum_{i \in Agv} w_{iv}^2 \, e_{iv}^2 \mid F_v \right\}, \quad (10)$$

where

$$\tilde{\theta}_{\text{FE}v} = \hat{\theta}_v + \sum_{g_v=1}^{Gv} \sum_{i \in A_{gv}} w_{iv}(\pi_{gv}^{-1} R_{iv} - 1) e_{iv}.$$

The estimator (7) can be implemented by using fractional imputation in which every responding unit in an imputation cell is used as a donor for every nonrespondent in the cell. Then, the estimator (7) can be written as the fractionally imputed estimator

$$\hat{\theta}_{\text{FEFI}} = \sum_{g=1}^{G} \sum_{j \in A \cap U_g} \sum_{i \in A_R \cap U_g} w_j w_{ij}^* y_i, \quad (11)$$

where w_j w_{ij}^* is the weight of donor i for recipient j, w_{ij}^* is the imputation fraction of donor i for recipient j defined in (3), and

$$w_{ij}^* = \begin{cases} \left(\sum_{s \in A_R \cap U_g} w_s \right)^{-1} w_i R_i & \text{if } R_j = 0 \\ 1 & \text{if } R_j = 1 \text{ and } i = j. \end{cases} \quad (12)$$

The estimator (11) with w_{ij}^* of (12), algebraically equivalent to (7), is called the *fully efficient fractionally imputed* (FEFI) estimator. The fractionally imputed estimator has the advantage that functions of y such as the fraction less than a given number can be directly estimated from the fractionally imputed data set.

To consider replication variance estimation, let a replication variance estimator for the complete sample be

$$\hat{V}(\hat{\theta}) = \sum_{k=1}^{L} c_k (\hat{\theta}^{(k)} - \hat{\theta})^2, \quad (13)$$

where $\hat{\theta}^{(k)}$ is the k^{th} estimate of θ_N based on the observations included in the k^{th} replicate, L is the number of replicates, and c_k is a factor associated with replicate k determined by the replication method. For a discussion of replication for survey samples see Krewski and Rao (1981) and Rao, Wu and Yue (1992). When the original estimator $\hat{\theta}$ is a linear estimator of the form (1), the k^{th} replicate estimate of $\hat{\theta}$ can be written

$$\hat{\theta}^{(k)} = \sum_{i \in A} w_i^{(k)} y_i, \quad (14)$$

where $w_i^{(k)}$ denotes the replicate weight for the i^{th} unit of the k^{th} replication.

A proposed replicate for the estimator $\hat{\theta}_{\text{FEFI}}$ is

$$\hat{\theta}_{\text{FEFI}}^{(k)} = \sum_{g=1}^{G} \left(\sum_{i \in A \cap U_g} w_i^{(k)} \right) \frac{\sum_{i \in A_R \cap U_g} w_i^{(k)} y_i}{\sum_{i \in A_R \cap U_g} w_i^{(k)}} =: \sum_{g=1}^{G} \sum_{j \in A \cap U_g} \sum_{i \in A_R \cap U_g} w_j^{(k)} w_{ij}^{*(k)} y_i. \quad (15)$$

Using the replicates (15), the replicate variance estimator can be written as

$$\hat{V}_{\text{FEFI}} = \sum_{k=1}^{L} c_k (\hat{\theta}_{\text{FEFI}}^{(k)} - \hat{\theta}_{\text{FEFI}})^2. \quad (16)$$

The replicates in (15) can be computed in two steps. First, create the usual replicate by defining the weights $w_i^{(k)}$ for every element. Second, for a nonrespondent, the replicate imputation fraction for donor i to recipient j is

$$w_{ij}^{*(k)} = \frac{w_i^{(k)}}{\sum_{s \in A_R \cap U_g} w_s^{(k)}}.$$

Note that the sum of the fractional replication weights of the donor records for each recipient is the same as the replication weight for that unit in a complete sample.

The suggested procedure is closely related to the Rao and Shao (1992) variance estimator. See also Yung and Rao (2000). However, the use of fractional imputation greatly simplifies variance estimation. In the creation of replicates, only the weights on the imputed values are changed. No recomputing of imputed values is required, and once computed, the replicate weights can be used for any smooth function of the vector y. Also, the fractional replicates make the estimator (16) appropriate for a vector of y–variables.

Theorem 3.1 of Kim, Navarro and Fuller (2005) can be used to show that, given a consistent full sample replication procedure,

$$\hat{V}_{\text{FEFI}} = V(\tilde{\theta}_{\text{FE}v} \mid F_v) - N_v^{-2} \sum_{g_v=1}^{G_v} \sum_{i \in U_{gv}} \pi_{gv}^{-1}(1-\pi_{gv}) e_{iv}^2 + o_p(n_v^{-1}), \quad (17)$$

where $\tilde{\theta}_{\text{FE}v}$ is defined in (10), and the distribution is with respect to the sampling and response mechanisms.

If the finite population correction can be ignored, the estimator (16) is consistent for $V\{\hat{\theta}_{\text{FE}}\}$. If the sample size is large relative to N, then an estimator of

$$N_v^{-2} \sum_{g_v=1}^{G_v} \sum_{i \in U_{gv}} \pi_{gv}^{-1}(1-\pi_{gv}) \, e_{iv}^2$$

should be added to (16).

The imputation and variance estimation procedure outlined for the response model also produces consistent estimators for the cell mean model. Under the cell mean model, the elements within a cell of the finite population are a realization of independently and identically distributed random variables. The imputation procedure based on the response model is not necessarily fully efficient for the population mean under the cell mean model, but it can be shown that the estimator of the mean and the estimator of the variance of the estimated mean are consistent.

4. Approximations to the Fully Efficient Procedure

In the previous sections, the estimator $\hat{\theta}_{\text{FEFI}}$ was constructed to produce zero imputation variance. The implementation of the fractional imputation procedure, as described in (11), could require the use of a large number of donors for each recipient. Therefore, we outline a procedure with a fixed number of donors per recipient that is fully efficient for the grand total, but not necessarily fully efficient for subpopulations. The procedure assigns donors to produce small between-recipient variance of imputed values and modifies the weights of donors to attain full efficiency for the total.

Suppose that M donors are to be assigned to each recipient. We suggest donors be assigned to recipients to approximate the distribution of all respondents in the cell. One possible selection method is to select a stratified sample for each recipient. A second possibility is to use systematic sampling with probability proportional to the weights to select donors for each recipient. Initial fractions w_{ij0}^* are assigned to the donated values. For systematic sampling with equal weights, the initial w_{ij0}^* is M^{-1}.

After the donors are assigned, the initial fractions, w_{ij0}^* are adjusted so that the sum of the weights gives the fully efficient estimator of the mean of y, and such that the estimated cumulative distribution function based on the weights approximates the fully efficient estimator of the cumulative distribution function. The modification of weights using regression has been suggested by Fuller (1984, 2003). Chen, Rao and Sitter (2000) discussed an efficient imputation method that changes the imputed values rather than the weights. Let $\mathbf{z}_{gj} = (z_{gj1}, z_{gj2}, \ldots, z_{gj\alpha})$ be a vector defined by

$$
\begin{aligned}
z_{gj1} &= y_j \\
z_{gj2} &= 1 \quad \text{if } y_j \le L_2 \\
&= 0 \quad \text{otherwise} \\
&\vdots \\
z_{gj\alpha} &= 1 \quad \text{if } L_{\alpha-1} < y_j \le L_\alpha \\
&= 0 \quad \text{otherwise,}
\end{aligned}
$$

where $L_2, L_3, \ldots, L_\alpha$ divide the range of observed y in cell g into $\alpha - 1$ sections. The number of sections that can be used depends on the numbers and type of observations in the cell, the number of recipients and the number of donors per recipient. If the number of donors per recipient is large, it is possible to adjust the set of weights for each recipient so that the sum of w_{ij}^* over i is one for every j and the sum of $w_{ij}^* y_i$ over i is the fully efficient estimator for every j. In most cases the weights will be adjusted so that the sum of the w_{ij}^* over i is one for every j and the cell means of the imputed values are equal to the fully efficient estimator.

Let $\bar{\mathbf{z}}_{\text{FE, g}}$ denote the fully efficient estimator for cell g. Using regression procedures, the modified w_{ij}^*, modified to give the fully efficient cell mean of $\mathbf{z}$, are

$$
w_{ij}^* = w_{ij0}^* + (\bar{\mathbf{z}}_{\text{FE},g} - \bar{\mathbf{z}}_g^*)\, \mathbf{S}_{\mathbf{zz}g}^{-1} w_{ij0}^* (\bar{\mathbf{z}}_{g[i]j} - \bar{\mathbf{z}}_{g\cdot j})', \qquad (18)
$$

where

$$
\mathbf{S}_{\mathbf{zz}g} = \sum_{j \in A_{Lg}} b_j \sum_{i \in A_{Rg}} w_{ij0}^* (\mathbf{z}_{g[i]j} - \bar{\mathbf{z}}_{g\cdot j})'(\mathbf{z}_{g[i]j} - \bar{\mathbf{z}}_{g\cdot j}) d_{ij},
$$

$$
\bar{\mathbf{z}}_{g\cdot j} = \sum_{i \in A_{Rg}} w_{ij0}^* \mathbf{z}_{g[i]j}\, d_{ij},
$$

$$
\bar{\mathbf{z}}_g^* = \sum_{j \in A_{Lg}} b_j \sum_{i \in A_{Rg}} w_{ij0}^* \mathbf{z}_{g[i]j}\, d_{ij},
$$

$$
b_j = \left(\sum_{s \in A_{Lg}} w_s \right)^{-1} w_j,
$$

A_{Lg} is the set of indexes of recipients in cell g, $\mathbf{z}_{g[i]j} = \mathbf{z}_{gi}$ is the value imputed from donor i for recipient j, and $\bar{\mathbf{z}}_{g\cdot j}$ is the weighted mean of the imputed values for recipient j using the initial w_{ij0}^*.

To estimate the variance, replicates are created so that the weights on the donors reflect the effect of the deletion of an element on the fully efficient estimator. We use the words "deletion" and "delete" to identify the element chosen for principal weight modification for replication variance estimation.

Let $w_i^{(k)}$ be the weight assigned to element i for the k^{th} replicate for variance estimation of the full sample estimator. Then the replicate for the fully efficient mean of y for cell g is

$$\bar{\mathbf{z}}_g^{(k)} = \left[\sum_{i \in A_{Rg}} w_i^{(k)} \right]^{-1} \sum_{i \in A_{Rg}} w_i^{(k)} \mathbf{z}_i. \tag{19}$$

Replicate fractions are assigned to donors in cell g so that the replicate estimate of the cell mean is $\bar{\mathbf{z}}_g^{(k)}$. Initial fractional weights $w_{ij0}^{*(k)}$ are assigned where $w_{ij0}^{*(k)}$ is small, but positive, if i is a deleted unit for replicate k. The final fractional weights $w_{ij}^{*(k)}$ are computed using the procedure of (18) with $\bar{\mathbf{z}}_g^{(k)}$ replacing $\bar{\mathbf{z}}_{\text{FE},g}$ and $w_{ij0}^{*(k)}$ replacing w_{ij0}^{*}. The procedure simulates the effect of deleting a single element on the fully efficient estimator.

5. An Artificial Example

In this section, we present an example with artificial data to illustrate the implementation of the proposed method. Suppose that two study variables, x and y, are observed in a sample of size $n = 10$ obtained by simple random sampling. Variable x is a categorical variable with three categories, say 1, 2, and 3, and variable y is a continuous variable. Both variables have item nonresponse and there is a set of imputation cells for each variable. Table 5.1 shows the sample observations, where nonresponse is denoted by M in the table. We use a weight of one to simplify the presentation. Divide by ten to obtain weights for the mean.

Table 5.1
An Illustrative Data Set

Observation	Weight	Cell for x	Cell for y	x	y
1	1	1	1	1	7
2	1	1	1	2	M
3	1	1	2	3	M
4	1	1	1	M	14
5	1	1	2	1	3
6	1	2	1	2	15
7	1	2	2	3	8
8	1	2	1	3	9
9	1	2	2	2	2
10	1	2	1	M	M

Because the x variable is a categorical variable with three categories, using three fractions for fractional imputation gives fully efficient estimators for the distribution of the $x-$variable. Thus the weights in Table 5.2 for the three imputed values of x for observation four are the fractions for the three categories in $x-$cell one.

If a subset of donors is to be used for each recipient, a controlled method of selecting donors, such as systematic sampling, is suggested. In our simple illustration we could easily use fractional imputation with all four y responses in cell 1, but to illustrate the regression adjustment we use only three. See Table 5.2.

Several approaches are possible for the situation in which two items are missing, including the definition of a third set of imputation cells for such cases. Because of the small size of our illustration, we impute under the assumption that x and y are independent within cells. Thus we impute four values for observation ten. For each of the two possible values of x we impute two possible values for y. One of the pair of imputed $y-$values is chosen to be less than the mean of responses and one is chosen to be greater than the mean. See the imputed values for observation 10 in Table 5.2.

Table 5.2
Fractional Weights for Means

Observation	Weight	Donor for y	Cell for x	Cell for y	x	y
1	1.0000		1	1	1	7
2	0.2886	1	1	1	2	7
2	0.3960	6	1	1	2	15
2	0.3154	8	1	1	2	9
3	0.3333	5	1	2	3	3
3	0.3333	7	1	2	3	8
3	0.3334	9	1	2	3	2
4	0.5000		1	1	1	14
4	0.2500		1	1	2	14
4	0.2500		1	1	3	14
5	1.0000		1	2	1	3
6	1.0000		2	1	2	15
7	1.0000		2	2	3	8
8	1.0000		2	1	3	9
9	1.0000		2	2	2	2
10	0.2247	8	2	1	2	9
10	0.2753	4	2	1	2	14
10	0.2095	1	2	1	3	7
10	0.2905	6	2	1	3	15

Initial fractions of one third are assigned to the three imputed values for observations three and four, and initial fractions of one fourth are assigned to the four imputed values for observation ten. The fractional weights are then adjusted using the regression method of equation (18) to give the FEFI mean of y as the estimator, where the fully efficient estimator for the mean of y is

$$\bar{y}_{\text{FE}} = \sum_{g=1}^{2} \frac{n_g}{n} \bar{y}_{Rg} = 8.4833.$$

We restrict the weights for observation 10 so that the estimated fractions for the two categories of x are the cell fractions. Then, because the weighted mean for the categorical variable is controlled for each individual, the vector $\mathbf{z}$ contains only the $y-$variable. Table 5.2 gives the final fractional weights computed with the regression weighting.

An analyst can use the data set of Table 5.2 and any full-sample computer program to compute estimates of functions of y and x, such as the mean of y for the x categories. The fractional data set is fully efficient for any function of the $x-$variable and is also fully efficient for the mean of the $y-$variable.

For jackknife variance estimation, we repeat the weight calculation for each replicate. The replicate estimates of the cell means of y are given in Table 5.3 and the replicate

estimates of the category fractions for x are given in Table 5.4. The values in Table 5.3 and in Table 5.4 are used as the control totals $\bar{\mathbf{z}}_{\mathrm{FE},g}$ in the regression weighting. We used $w_{ij0}^{*(k)} = 3^{-1}$ as the initial value of the replication fractions for observation two and $w_{ij0}^{*(k)} = 4^{-1}$ for observation ten.

Table 5.5 contains the jackknife weights for the fractionally imputed data set of Table 5.2. The replicate weights are used in the same way as replicates for a full sample. They are appropriate, with the caveats of the next section, for any statistic for which the full sample jackknife is appropriate. Thus the procedure is particularly appealing for a general purpose data set, because no additional computations are required of the analyst.

The fully efficient estimator of the mean of y is obtained by treating the respondents as the second phase of a two phase sample. A two-phase variance estimator is

$$\hat{V} = \frac{1}{n}\sum_{g=1}^{2}\frac{n_g}{n}(\bar{y}_{Rg} - \bar{y}_{\mathrm{FE}})^2 + \sum_{g=1}^{2}\left(\frac{n_g}{n}\right)^2\frac{1}{r_g}s_{Rg}^2 = 3.043,$$

where s_{Rg}^2 is the within cell sample variance for cell g. If we use the replication weights in Table 5.5, the replication variance estimate for the mean of y is

$$\hat{V}_{\mathrm{JK}}(\bar{y}_{\mathrm{FI}}) = \sum_{k=1}^{10} 0.9\,(\bar{y}_{\mathrm{FI}}^{(k)} - \bar{y}_{\mathrm{FI}})^2 = 3.078.$$

The difference between the linearized variance estimator and the jackknife variance estimator is

$$\sum_{g=1}^{2}\left(\frac{r_g}{r_g - 1}\frac{n-1}{n} - 1\right)s_{Rg}^2.$$

Thus, the jackknife variance estimator slightly overestimates the true variance in this example.

Table 5.3
Jackknife Replicates of Cell Mean of $y-$variable

Cell	Replicate									
	1	2	3	4	5	6	7	8	9	10
1	12.67	11.25	11.25	10.33	11.25	10.00	11.25	12.00	11.25	11.25
2	4.33	4.33	4.33	4.33	5.00	4.33	2.50	4.33	5.50	4.33

Table 5.4
Jackknife Replicates of Cell Mean of the Dummy Variables of $x-$variable

Cell	Level of x	Replicate									
		1	2	3	4	5	6	7	8	9	10
	1	0.33	0.67	0.67	0.50	0.33	0.50	0.50	0.50	0.50	0.50
1	2	0.33	0.00	0.33	0.25	0.33	0.25	0.25	0.25	0.25	0.25
	3	0.33	0.33	0.00	0.25	0.33	0.25	0.25	0.25	0.25	0.25
	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	2	0.50	0.50	0.50	0.50	0.50	0.33	0.67	0.67	0.33	0.50
	3	0.50	0.50	0.50	0.50	0.50	0.67	0.33	0.33	0.67	0.50

Table 5.5
Jackknife Weights for Fractional Imputation

Obs.	Replicate									
	1	2	3	4	5	6	7	8	9	10
1	0	1.1111	1.1111	1.1111	1.1111	1.1111	1.1111	1.1111	1.1111	1.1111
2	0.1664	0	0.3206	0.4205	0.3206	0.4563	0.3206	0.2392	0.3206	0.2724
2	0.6559	0	0.4400	0.3002	0.4400	0.2500	0.4400	0.5540	0.4400	0.5075
2	0.2888	0	0.3505	0.3904	0.3505	0.4048	0.3505	0.3179	0.3505	0.3312
3	0.3706	0.3706	0	0.3706	0.3226	0.3706	0.5018	0.3706	0.2867	0.3706
3	0.3697	0.3697	0	0.3697	0.5018	0.3697	0.0090	0.3697	0.6004	0.3697
3	0.3708	0.3708	0	0.3708	0.2867	0.3708	0.6003	0.3708	0.2240	0.3708
4	0.3703	0.7407	0.7407	0	0.3703	0.5556	0.5556	0.5556	0.5556	0.5556
4	0.3704	0	0.3704	0	0.3704	0.2777	0.2777	0.2777	0.2777	0.2777
4	0.3704	0.3704	0	0	0.3704	0.2778	0.2778	0.2778	0.2778	0.2778
5	1.1111	1.1111	1.1111	1.1111	0	1.1111	1.1111	1.1111	1.1111	1.1111
6	1.1111	1.1111	1.1111	1.1111	1.1111	0	1.1111	1.1111	1.1111	1.1111
7	1.1111	1.1111	1.1111	1.1111	1.1111	1.1111	0	1.1111	1.1111	1.1111
8	1.1111	1.1111	1.1111	1.1111	1.1111	1.1111	1.1111	0	1.1111	1.1111
9	1.1111	1.1111	1.1111	1.1111	1.1111	1.1111	1.1111	1.1111	0	1.1111
10	0.1624	0.2777	0.2777	0.3061	0.2777	0.2286	0.3474	0.3013	0.1520	0
10	0.3931	0.2778	0.2778	0.2494	0.2778	0.1417	0.3934	0.4395	0.2185	0
10	0.0932	0.2778	0.2778	0.3231	0.2778	0.4400	0.1483	0.0746	0.3171	0
10	0.4623	0.2778	0.2778	0.2324	0.2778	0.3008	0.2220	0.2957	0.4235	0

6. Simulation Studies

6.1 Study Parameters

To study the properties of the imputation procedure we conducted a Monte Carlo study. The sample is a stratified sample with two elements per stratum and two imputation cells, where the cells cut across the strata. Cell one is 20% of the population in strata 1–25 and 80% of the population in strata 26–50. The probability of response is 0.7 for cell one and 0.5 for cell two. Two variables are considered. The variable D is always observed and defines a subpopulation. The probability that $D = 1$ is 0.25 for cell one and 0.40 for cell two. The variable y is subject to nonresponse with constant within-cell response probabilities. The variable D is independent of y and of the response probability. The variable y is normally distributed, where the parameters for a population of 50 strata are given in Table 5.1. In the data generating model of Table 6.1, there are no stratum effects. The parameters of interest are: $\theta_1 =$ mean of y, $\theta_2 =$ mean of y for $D = 1$, $\theta_3 =$ fraction of Y's less than two, $\theta_4 =$ fraction of Y's less than one.

Table 6.1
Parameter Set A

		Cell One		Cell Two	
Strata	Element Weight	Mean	Variance	Mean	Variance
1–25	0.01	0.4	0.36	1.6	0.36
26–50	0.01	0.4	0.36	1.6	0.36

6.2 Estimation Procedures

In the simulation $M = 5$ and $M = 3$ donors were used per recipient. Systematic samples were selected to serve as donors for each recipient. If the number of respondents in the cell is less then M, every respondent was used as a donor for every recipient and the w_{ij}^* are proportional to the original w_i of the respondents. If there are more than M respondents in a cell, the donors are ordered by size and numbered from one to r_g. Then the donors are placed in the order $1, 3, 5, \ldots, r_g, r_{g-1}, r_{g-3}, \ldots, 2$ for r_g odd and the order $1, 3, 5, \ldots, r_{g-1}, r_g, r_{g-2}, \ldots, 2$ for r_g even. The cumulated sums of the weights are formed and m_g systematic samples of size M are selected, where $m_g = n_g - r_g$. The cumulative sums are normalized so that the grand sum is one, a random number, R_{Ng}, between zero and $0.2m_g$ is selected and the m_g samples are the systematic samples of size M defined by the donor associated with $R_{Ng} + 0.2(s-1) + (t-1)m_g^{-1}$, $s = 1, 2, 3, 4, 5$ for recipients $t = 1, 2, \ldots, m_g$. The initial imputation fraction for each donor is $w_{ij}^* = M^{-1}$.

The initial imputation fractions are modified using the regression procedure of (18). The donors in a cell were ordered from smallest to largest and the cumulative sum of the weights formed. Let

$$S_{g,wt} = \sum_{i=1}^{t} w_{[i]}, i \in A_{Rg}, \tag{20}$$

where $w_{[i]}, i = 1, 2, \ldots, r_g$, is the weight of $y_{g,(i)}$ and $y_{g,(1)} \le \ldots \le y_{g,(n)}$ are the ordered y–values in cell g. To define the boundaries of groups to be used to create indicator functions, let t_{*s} be the t for which

$$\max \{S_{g,wt} : S_{g,wt} \le 0.2sS_{gw}\}$$

for $s = 1, 2, 3, 4$, where S_{gw} is the total of the weights of the donors in cell g. Define

$$\begin{aligned} z_{gi,s+1} &= 1 \quad \text{if } y_i \le y_{g,(t_{*s})} \text{ and } i \in A_{Rg} \\ &= 0 \quad \text{otherwise} \end{aligned} \tag{21}$$

for $s = 1, 2, 3, 4$ and let $\mathbf{z}_{gj} = (y_{gj1}, z_{gj2}, \ldots, z_{gj5})$. The regression modified imputed estimator of the mean for each of the five variables in the z–vector is the fully efficient estimator of the respective mean.

The k–deleted FE estimator of the cell mean of $\mathbf{z}$ is defined in (19). The initial fractional weight for donor k to element j is set at $w_{kj0}^{*(k)} = 0.01 w_{kj}^*$. This initial weight assures that the final weight will be small, but permits regression adjustment. The final $w_{ij}^{*(k)}$ are computed using the regression procedure of (18) using the initial weight $w_{ij0}^{*(k)}$.

6.3 Monte Carlo Results

The Monte Carlo results for 5,000 samples generated by the parameters of Table 6.1 are given in Table 6.2 and Table 6.3. Results are given for the full sample, for fractional imputation with 5 donors, fractional imputation with three donors, and for multiple imputation (MI) using the Approximate Bayesian Bootstrap (ABB) of Rubin and Schenker (1986) with $M = 5$ and ABB with $M = 3$. Both the FI and MI procedures are unbiased for all four parameters of Table 6.2. The last column of Table 6.2 gives the Monte Carlo variance of the estimator divided by the Monte Carlo variance of the FI procedure with $M = 5$, expressed in percent. The FI procedure is five to ten percent more efficient than MI with $M = 5$ and 9 to 13 percent more efficient than MI with $M = 3$.

Under the model, the mean of the observed values is not the best estimator of the domain mean. In this example, the FI estimator is about as efficient as the full sample estimator. The effect of a smaller number of observations is balanced by the use of a superior estimator of the mean for the domain. Under the model, the domain indicator is independent of the y values, given the cell. Therefore it is efficient to use all values in the cell as donors, not just respondents in the domain.

The properties of the variance estimators are given in Table 6.3. The column headed "Relative Mean" gives the Monte Carlo estimated mean of the estimated variances

divided by the Monte Carlo estimated variance, where the Monte Carlo estimated variance is given in Table 6.2. Both variance estimation procedures appear to be nearly unbiased for the variance of the mean. The relative variance of the MI variance estimator for $M = 5$ is nearly twice that of the FI variance estimator for $M = 5$. For $M = 3$, the MI variance estimator is more than three times that for FI. The MI variance estimator has a large variance because the variance due to missing observations is estimated with four degrees-of-freedom for $M = 5$ and with two-degrees-of freedom for $M = 3$.

The MI variance estimator for the domain mean is seriously biased. This property was first identified by Fay (1991, 1992) and studied by Meng (1994) and Wang and Robins (1998). The FI variance estimator for the domain mean also has a positive bias, though much smaller than that of MI. The bias in the FI variance estimator can be reduced by increasing M, but the bias of MI has little relationship to M.

All variance estimators for the variance of $\hat{\theta}_4$ are slightly negatively biased. We believe FI is slightly biased for $\hat{\theta}_4$ because, although we use the z–vector, the weights are slightly smoothed by the regression procedure. MI is known to have a small sample bias. See Kim (2002).

Table 6.2
Mean and Variance of the Point Estimators Under Setup A (5,000 Samples of Size 100)

Parameter	Imputation Scheme	Mean	Variance	Stand. Var.
Mean	Complete Sample	1.00	0.00570	67
(θ_1)	FI(3)	1.00	0.00849	100
	ABB(3)	1.00	0.00926	109
	FI(5)	1.00	0.00849	100
	ABB(5)	1.00	0.00903	106
Domain Mean	Complete Sample	1.14	0.02020	99
(θ_2)	FI(3)	1.14	0.02050	100
	ABB(3)	1.14	0.02230	109
	FI(5)	1.14	0.02040	100
	ABB(5)	1.14	0.02170	106
$\Pr(Y < 2)$	Complete Sample	0.87	0.00104	51
(θ_3)	FI(3)	0.87	0.00202	100
	ABB(3)	0.87	0.00228	113
	FI(5)	0.87	0.00202	100
	ABB(5)	0.87	0.00223	110
$\Pr(Y < 1)$	Complete Sample	0.50	0.00208	66
(θ_4)	FI(3)	0.50	0.00313	100
	ABB(3)	0.50	0.00342	109
	FI(5)	0.50	0.00313	100
	ABB(5)	0.50	0.00329	105

Table 6.3
Relative Mean, t–statistic and Relative Variance for the Variance Estimators Under Setup A (5,000 Samples of Size 100)

Parameter	Method	Relative Mean (%)**	t–statistic*	Relative Variance (%)
Mean	FI(3)	100.1	0.05	5.66
(θ_1)	ABB(3)	99.6	–0.19	19.25
	FI(5)	100.1	0.03	5.65
	ABB(5)	98.2	–0.89	9.95
Domain Mean	FI(3)	115.9	7.54	13.88
(θ_2)	ABB(3)	127.9	12.72	28.88
	FI(5)	106.6	3.14	11.62
	ABB(5)	128.4	13.43	20.03
$\Pr(Y < 2)$	FI(3)	103.9	1.86	13.90
(θ_3)	ABB(3)	100.8	0.36	48.42
	FI(5)	101.7	0.82	12.07
	ABB(5)	98.5	–0.67	25.10
$\Pr(Y < 1)$	FI(3)	98.5	–0.75	4.67
(θ_4)	ABB(3)	96.3	–1.80	18.51
	FI(5)	97.6	–1.20	4.45
	ABB(5)	96.7	–1.65	10.17

* Statistic for hypothesis that the estimated variance is unbiased.
** Monte Carlo mean of variance estimates divided by Monte Carlo variance of estimates, in percent.

In a second set of parameters, denoted by C, the means were as follows:

$$\begin{aligned}&\text{Cell 1 of strata } 1-25; \quad \mu = 0.4\\ &\text{Cell 1 of strata } 26-50; \quad \mu = 3.0\\ &\text{Cell 2 of strata } 1-25; \quad \mu = 1.6\\ &\text{Cell 2 of strata } 26-50; \quad \mu = 2.2.\end{aligned}$$

All other parameters are the same as in parameter set A. The properties of the estimators are given in Table 6.4. Both FI and MI produce unbiased estimates of the means and of the domain mean. As with parameter set A, the FI procedure is eight to twelve percent more efficient than MI for $M = 5$ and 14 to 16 percent more efficient for $M = 3$.

The assumptions required for MI variance estimation are not satisfied for parameter set C. Therefore the MI estimated variance is seriously biased for all parameters. See Table 6.5. The bias in the MI estimated variance with $M = 5$ is about 17% for the variance of the overall mean and nearly 50% for the domain mean. The bias of the MI variance of the mean for a binomial variable is smaller than the bias for the mean of the continuous variable because the stratification effect is smaller for the binomial variable.

The properties of the estimated variances for the FI procedures are similar to those for setup A. There is a positive bias for the variance of the domain mean of about 23% for $M = 3$ and about 6% for $M = 5$.

The variance of the MI estimated variance is 2.4 to 3.5 times the variance of the FI estimated variance for $M = 5$ and 3 to 7 times for $M = 3$, demonstrating the clear superiority of the FI variance estimator for this configuration.

Table 6.4
Mean and Variance of the Point Estimators Under Setup C (5,000 Samples of Size 100)

Parameter	Imputation Scheme	Mean	Variance	Stand.Variance
Mean	Complete Sample	2.10	0.00500	48
(θ_1)	FI(3)	2.10	0.01050	100
	ABB(3)	2.10	0.01220	116
	FI(5)	2.10	0.01050	100
	ABB(5)	2.10	0.01150	110
Domain Mean	Complete Sample		0.02530	102
(θ_2)	FI(3)	2.01	0.02510	101
	ABB(3)	2.01	0.02850	115
	FI(5)	2.01	0.02480	100
	ABB(5)	2.01	0.02710	109
$\Pr(Y < 2)$	Complete Sample		0.00127	45
(θ_3)	FI(3)	0.45	0.00281	100
	ABB(3)	0.45	0.00322	115
	FI(5)	0.45	0.00280	100
	ABB(5)	0.45	0.00314	112
$\Pr(Y < 1)$	Complete Sample		0.00107	54
(θ_4)	FI(3)	0.15	0.00199	100
	ABB(3)	0.15	0.00226	114
	FI(5)	0.15	0.00199	100
	ABB(5)	0.15	0.00214	108

Table 6.5
Relative Mean, t–statistic and Relative Variance for the Variance Estimators Under Setup C (5,000 Samples of Size 100)

Parameter	Method	Relative Mean (%)	t–statistic*	Relative Variance (%)
Mean	FI(3)	100.9	0.41	6.42
(θ_1)	ABB(3)	116.7	7.31	40.14
	FI(5)	100.8	0.39	6.42
	ABB(5)	117.1	7.99	22.29
Domain Mean	FI(3)	122.7	10.78	16.23
(θ_2)	ABB(3)	144.4	19.79	46.05
	FI(5)	106.1	2.95	11.95
	ABB(5)	148.7	22.51	32.49
$\Pr(Y < 2)$	FI(3)	104.4	2.18	6.63
(θ_3)	ABB(3)	114.7	6.54	42.32
	FI(5)	101.8	0.89	6.42
	ABB(5)	112.1	5.74	20.67
$\Pr(Y < 1)$	FI(3)	102.3	1.13	11.08
(θ_4)	ABB(3)	101.3	0.58	39.14
	FI(5)	99.9	–0.04	10.05
	ABB(5)	102.2	1.04	23.60

* Statistic for hypothesis that the estimated variance is unbiased.

7. Summary

In fractional imputation, several donors are used for each missing value and each donor is given a fraction of the weight of the nonrespondent. If all donors are used, the procedure is fully efficient, under the model, for all functions of a **y**–vector. It is shown that the use of fractional imputation with a small number of imputations per nonrespondent can give a fully efficient estimator of the mean. Estimates of other parameters, such as estimates of the cumulative distribution are nearly fully efficient.

Fractional imputation permits the construction of general purpose replicates for variance estimation. A single set of replicates can be used for variance estimation for imputed variables, variables observed on all respondents, and under model assumptions, for functions of the two types of variables. The replicates give estimates of the variances of domain means with much smaller biases than those of multiple imputation. The bias goes to zero as M increases and, in the simulation, is modest for $M = 5$. The replication variance estimator is easily implemented with replication software such as Wesvar.

Fractional imputation with a fixed number of donors per recipient is slightly more efficient for the mean than multiple imputation with the same number of donors. Fractional imputation gives variance estimates with smaller bias and much smaller variance than multiple imputation estimators with the same number of imputations.

8. Acknowledgements

This research was partially supported by a subcontract between Westat and Iowa State University under Contract No. ED-99-CO-0109 between Westat and the Department of Education and by Cooperative Agreement 13-3AEU-0-80064 between Iowa State University, the U.S. National Agricultural Statistics Service and the U.S. Bureau of the Census. We thank Jean Opsomer and Damiao Da Silva for useful comments.

References

Chen, J., Rao, J.N.K. and Sitter, R.R. (2000). Efficient random imputation for missing data in complex surveys. *Statistica Sinica*, 10, 1153-1169.

Fay, R.E. (1991). A design-based perspective on missing data variance. *Proceedings of Bureau of the Census Annual Research Conference*, American Statistical Association, 429-440.

Fay, R.E. (1992). When are inferences from multiple imputation valid? *Proceedings of the Survey Research Methods Section*, American Statistical Association, 227-232.

Fay, R.E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association*, 91, 490-498.

Ford, B.M. (1983). An overview of hot-deck procedures. In *Incomplete Data in Sample Surveys*, Vol. 2. New York: Academic Press, 185-207.

Fuller, W.A. (1984). Least squares and related analyses for complex survey designs. *Survey Methodology*, 10, 97-118.

Fuller, W.A. (1996). *Introduction to Statistical Time Series*. 2nd Edition. New York: John Wiley & Sons, Inc.

Fuller, W.A. (2003). Estimation for multiple phase samples. In *Analysis of Survey Data*, (Eds. R.L. Chambers and C.J. Shinner). Wiley, Chichister, England, 307-322.

Kalton, G., and Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1-16.

Kalton, G., and Kish, L. (1984). Some efficient random imputation methods. *Communications in Statistics Part A – Theory and Methods*, 13, 1919-1939.

Kim, J.K. (2002). A note on approximate Bayesian bootstrap imputation. *Biometrika*, 89, 470-477.

Kim, J.K., and Fuller, W.A. (2004). Fractional hot deck imputation. *Biometrika*, 91, 559-578.

Kim, J.K., Navarro, A. and Fuller, W.A. (2005). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Association*, to appear.

Kott, P.S., and Stukel, D.M. (1997). Can the jackknife be used with a two-phase sample? *Survey Methodology*, 23, 81-89.

Krewski, D., and Rao, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *The Annals of Statistics*, 9, 1010-1019.

Little, R.J.A., and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. 2nd Edition. New York: John Wiley & Sons, Inc.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 9, 538-573.

Rao, J.N.K. (1996). On variance estimation with imputed survey data. *Journal of the American Statistical Association*, 91, 499-506.

Rao, J.N.K., and Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.

Rao, J.N.K., and Sitter, R.R. (1995). Variance estimation under two-phase sampling with applications to imputation for missing data *Biometrika*, 82, 453-460.

Rao, J.N.K., Wu, C.F.J. and Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology*, 18, 209-217.

Rubin, D.B., and Schenker (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81, 366-374.

Rubin, D.B. (1987). *Multiple Imputation For Nonresponse In Surveys*. New York: John Wiley & Sons, Inc.

Sande, I.G. (1983). Hot-deck imputation procedures. *Incomplete Data in Sample Surveys*, Vol. 3. New York: Academic Press, 339-349.

Särndal, C.-E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, 18, 241-252.

Shao, J., Chen, Y. and Chen, Y. (1998). Balanced repeated replication for stratified multistage survey data under imputation. *Journal of the American Statistical Association*, 93, 819-831.

Shao, J., and Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94, 254-265.

Sitter, R.R. (1997). Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association*, 92, 780-787.

Tollefson, M., and Fuller, W.A. (1992). Variance estimation for sampling with random imputation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 140-145.

Wang, N., and Robins, J.M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika*, 85, 935-948.

Yung, W., and Rao, J.N.K. (2000). Jackknife variance estimation under imputation for estimators using poststratification information. *Journal of American Statistical Association*, 95, 903-915.